Sample Size and Statistical Power

Outline

1. Concepts associated with power analysis
2. Comparing means
3. Comparing proportions
4. Time to event studies
5. Longitudinal studies: continuous response
6. Longitudinal studies: binary response
7. Cluster or group randomized trials

1. Concepts associated with power analysis

Null hypothesis: $H_0: \theta = \theta_0$
1-sided alternative: $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$
2-sided alternative: $H_1: \theta \neq \theta_0$

Sample size calculations require specification of:

1. Significance level $\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$
2. Power: $1 - \beta = P(\text{reject } H_0 | H_1 \text{ is true})$
3. Smallest difference of clinical/biologic importance, d
4. Measurement variation, σ^2

May also need:

5. Number of repeats per person, m
6. Correlation between repeats, ρ

2. Comparing means

$$H_0\text{: } \mu_1 - \mu_2 = 0$$
$$H_1\text{: } \mu_1 - \mu_2 = d$$

The sample size per group is

$$n = \frac{2\sigma^2\left(z_{\alpha/2} + z_\beta\right)^2}{d^2}$$

For unequal groups of size n_1 and n_2, where $r = n_2/n_1$, is

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/r)\left(z_{\alpha/2} + z_\beta\right)^2}{d^2}$$

and $n_2 = rn_1$.

2. Comparing means

σ consists of two components: (i) inherent biological variation between subjects, σ_b, and (ii) measurement error, σ_e, where $\sigma^2 = \sigma_b^2 + \sigma_e^2$.

Ordinarily one hopes that σ_e^2 is small relative to σ_b^2. Poor measurement can inflate σ^2, leading to larger sample sizes. One can decrease σ^2 by reducing σ_e^2 through improved measurement, or by taking multiple measurements on each subject.

If one uses the mean of K replicates as the datum for each subject, $\sigma^2 = \sigma_b^2 + \sigma_e^2/K$.

3. Comparing proportions

Fleiss JL *Statistical Methods for Rates and Proportions* (2nd edition). Wiley:New York, 1981.

Let p_i be the proportion of subjects in group i having the outcome of interest, $\overline{p} = (p_1 + p_2)/2$ and $\overline{q} = 1 - \overline{p}$.

$$H_0\text{: } p_1 - p_2 = 0$$
$$H_1\text{: } p_1 - p_2 = d$$

The sample size per group is

$$n' = \frac{\left\{ z_{\alpha/2}\sqrt{2\overline{p}\,\overline{q}} + z_{\beta}\sqrt{p_1 q_1 + p_2 q_2} \right\}^2}{d^2}$$

$$n = n'/4\left(1 + \sqrt{1 + 4/n'|d|}\right)^2$$ "continuity correction".

3. Comparing proportions

For unequal groups of size n_1 and n_2, where $r = n_2/n_1$, is

$$n_1' = \frac{\left\{z_{\alpha/2}\sqrt{(r+1)\overline{p}\,\overline{q}} + z_{\beta}\sqrt{rp_1q_1 + p_2q_2}\right\}^2}{rd^2}$$

where $\overline{p} = \dfrac{p_1 + rp_2}{r+1}$ and $n_2 = rn_1$.

For small samples, employ a "continuity correction"

$$n_1 = \frac{n_1'}{4}\left(1 + \sqrt{1 + \frac{2(r+1)}{n_1' r|d|}}\right)^2.$$

3. Comparing proportions

Example: Want to compare rates of prematurity in infants born to women who attend prenatal clinics (hypothesized to be $p_1 = 0.25$) with non-attenders (hypothesized to be $p_2 = 0.40$). Because recruitment of non-attenders is difficult, we decide to study half as many non-attenders, or $r = n_2/n_1 = 0.5$.

Find n_1 and n_2 needed to ensure $\alpha = .01$ and 95% power.

Note: OR $= \dfrac{p_1 q_2}{p_2 q_1} = \dfrac{.25(1-.40)}{.40(1-.25)} = 0.50$

3. Comparing proportions

$$\overline{p} = \frac{p_1 + rp_2}{r+1} = \frac{.25 + .5(.40)}{1.5} = 0.30$$

$$n_1' = \frac{\left[2.576\sqrt{1.5(.3)(.7)} + 1.645\sqrt{.5(.25)(.75) + .4(.6)}\right]^2}{0.5(.15)^2}$$

$$= 510.34$$

$$n_1 = \frac{510.34}{4}\left(1 + \sqrt{1 + \frac{2(1.5)}{510.34(0.5)(0.15)}}\right)^2 = 530$$

$$n_2 = 0.5(530) = 265$$

$n_1 + n_2 = 795.$

4. Time to event studies

Freedman L.S. (1982) Tables of the number of patients required in clinical trials using the logrank test. *Stat in Med*, 1:121-129.

Lee E.T. (1992) *Statistical methods for survival analysis*. pg 340.

If all subjects are followed for the same fixed amount of time (say 3 years) we could simply compare the proportions of patients with the outcome of interest...and sample size formulas to compare two independent proportions apply.

4. Time to event studies

Few studies have equal follow-up times for all subjects, so survival analysis methods are necessary.

In randomized trials the usual statistic for comparing two survival functions is the non-parametric log-rank test. It does not make any assumptions about the survival distributions in the groups.

Even when the log-rank test is intended for analysis, often assumptions about the survival functions are made when determining the appropriate sample size.

4. Time to event studies

To compute sample size, the following is needed:

1. the estimated proportions of subjects in each group who are "endpoint-free" at a fixed time

2. or, the estimated hazard ratio ($h = e^{\beta}$ where β is the Cox model coefficient corresponding to the treatment effect), and estimated control group probability of survival at a fixed time.

3. or, the estimated median survival times or exponential hazard rates in each group.

4. Time to event studies

Based on log-rank test

This calculation assumes patients are followed for fixed length of time (t), and the hazard ratio (h) is constant over time. If p_i denotes the proportion of subjects who are endpoint-free at time t for group i, then

$$h = \frac{\ln(p_1)}{\ln(p_2)}$$

and the sample size per group is

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 (h+1)^2}{(2 - p_1 - p_2)(h-1)^2}$$

4. Time to event studies

Based on log-rank test

Example: Want to determine if a new drug for treatment of lung cancer lengthens survival time. All patients in the trial will be followed for 2 years. Find sample size needed for an $\alpha = 0.05$ level test to have 90% power to detect $h = 1.5$ if the 2-year survival rate under standard therapy is $p_1 = 0.25$.

$$h = 1.5 = \frac{\ln(.25)}{\ln(p_2)} \qquad \Rightarrow \qquad p_2 = 0.397.$$

$$n = \frac{(1.96 + 1.28)^2 (1.5 + 1)^2}{(2 - .25 - .397)(1.5 - 1)^2} = 195.$$

4. Time to event studies

Based on exponential survival and accrual

This calculation assumes patients enter the study at a uniform rate until the trial ends in t years, and survival curves are exponential with parameter λ_i, the hazard for group i (instanstaneous risk of failure)

The sample size per group is

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 [\phi(\lambda_1) + \phi(\lambda_2)]}{(\lambda_1 - \lambda_2)^2}$$

where

$$\phi(\lambda) = \frac{\lambda^3 t}{\lambda t + e^{-\lambda t} - 1}$$

4. Time to event studies

Based on exponential survival and accrual

If the survival curves are exponential, the hazard for group i can be estimated using

$$\lambda_i = -\frac{1}{t_{50}}\ln(0.5)$$

where t_{50} is the median survival time for group i.

Or, if you have λ for the control group, and h use

$$h = \frac{\lambda_1}{\lambda_2}$$

to determine λ_2.

4. Time to event studies

Based on exponential survival and accrual

Example: In the previous example, suppose the trial lasts 2 years and the survival curves are exponential. The median survival with the standard drug is 1 year, corresponding to $\lambda_1 = -\frac{1}{1}\ln(0.5) = 0.693$.

$$h = 1.5 = .693/\lambda_2 \qquad \Rightarrow \qquad \lambda_2 = 0.462.$$

$$\phi(\lambda_1) = \frac{(0.693)^3 \times 2}{(0.693)2 + e^{-.693\times 2} - 1} = 1.046; \quad \phi(\lambda_2) = 0.615$$

$$n = \frac{(1.96 + 1.28)^2[1.046 + 0.615]}{(0.693 - 0.462)^2} = 324$$

4. Time to event studies

The above calculations consider only the survival distributions, but not factors such as loss to follow-up, and shorter accrual period. These factors can be very important because they determine the number of failures expected in the trial, which in turn affects power. To take these factors into account, one should turn to commercial software packages.

5. Longitudinal studies: continuous response

Want to compare the rate of change over time in a continuous response. Let Y_{ijk} denote the response for subject i $(i = 1, ..., n)$ at time x_j $(j = 1, ..., m)$ within group k $(k = 1, 2)$. The model for computing sample size is

$$Y_{ijk} = \beta_{0k} + \beta_{1k} x_j + e_{ijk}$$

This model assumes each person is measured at same set times, $x_{ijk} = x_j$ for all i, k. If x_j is the time (in years) between the first and j-th visit, β_{1k} represent the annual "rate of change" in Y.

$$H_0 \text{: } \beta_{12} - \beta_{11} = 0$$
$$H_1 \text{: } \beta_{12} - \beta_{11} = d$$

5. Longitudinal studies: continuous response

Repeated measurements from a subject are correlated. For simplicity in sample size calculations, assume a simple covariance structure (compound symmetry):

$$V(Y_{ijk}) = \sigma^2 \text{ and } \text{corr}(Y_{ijk}, Y_{ij'k}) = \rho \text{ for } j \neq j'$$

The number of subjects per group is

$$n = \frac{2\sigma^2(1-\rho)\left(z_{\alpha/2} + z_{\beta}\right)^2}{m s_x^2 d^2} \text{ where } s_x^2 = \sum_{j=1}^{m}(x_j - \overline{x})^2/m$$

*Note that the sample size decreases as ρ increases.

5. Longitudinal studies: continuous response

Example: Want to test effectiveness of new drug in lowering blood pressure. Three visits are planned at years 0 (baseline), 2 and 5.

$$s_x^2 = \frac{(0-2.33)^2 + (2-2.33)^2 + (5-2.33)^2}{3} = 4.22$$

How large a sample is needed for an $\alpha = 0.05$ level test to have 80% power to detect an annual reduction in blood pressure that is 0.5 mmHg/year greater with the new drug? Assume $\sigma^2 = 100$.

$n = 313$ if $\rho = 0.2$
$n = 195$ if $\rho = 0.5$
$n = 79$ if $\rho = 0.8$

5. Longitudinal studies: continuous response

This calculation assumes linearity as a simplication for planning. Nevertheless, one should be prepared for a nonlinear response and propose statistical analyses to look for this.

Values for σ^2 and ρ have to be estimated from previous data or assumed. If longitudinal data are not available, one can use slope values based on cross-sectional regression of age on Y. For example, $\beta = -.476$ implies an average cross-sectional decline in Y of $.476$ units per year.

Note that we have not accounted for missing data, due to missed visits or dropout.

5. Longitudinal studies: continuous response

Suppose that instead of comparing slopes we want to compare average response for 2 groups. The model is

$$Y_{ij} = \beta_0 + \beta_1 x + e_{ij}, \quad i = 1, ..., 2n; \quad j = 1, ...m$$

where x is the treatment assignment indicator variable.

$$H_0 \text{: } \beta_1 = 0$$
$$H_1 \text{: } \beta_1 = d$$

The number of subjects per group is

$$n = \frac{2\sigma^2[1-(m-1)\rho]\left(z_{\alpha/2} + z_{\beta}\right)^2}{md^2}$$

*Note in this case, n <u>increases</u> as ρ increases.

5. Longitudinal studies: continuous response

Example: In previous example, find n needed for an $\alpha = 0.05$ level test to have 80% power to detect a 2 mmHg reduction in average blood pressure over 3 visits with the new drug. Assume $\sigma^2 = 100$.

$$n = 146 \text{ if } \rho = 0.2$$
$$n = 208 \text{ if } \rho = 0.5$$
$$n = 270 \text{ if } \rho = 0.8$$

6. Longitudinal studies: binary response

In this case, Y_{ijk} denotes the repeated binary response for subject i at time x_j within group k, with probability of success $(y_{ijk} = 1)$ being p_i for group i.

$$H_0 \colon p_1 - p_2 = 0$$
$$H_1 \colon p_1 - p_2 = d$$

The number of subjects needed per group is

$$n = \frac{\left\{ z_{\alpha/2} \sqrt{2\overline{p}\,\overline{q}[1 + (m-1)\rho]} + z_{\beta} \sqrt{(p_1 q_1 + \ p_2 q_2)[1 + (m-1)\rho]} \right\}^2}{md^2}$$

6. Longitudinal studies: binary response

Example: Find n needed for an $\alpha = 0.05$ level test to have 80% power to detect a 20% difference in the average probability of success over 3 repeated visits, for control vs. treated groups.

$$n = 35 \text{ if } \rho = 0.2$$
$$n = 49 \text{ if } \rho = 0.5$$
$$n = 64 \text{ if } \rho = 0.8$$

*As in the analogous problem with a continuous response, n increases as ρ increases.

7. Cluster or Group Randomized Trials

Murray D. *Design and Analysis of Group-Randomized Trials*. New York:Oxford University Press, 1998.

Group-randomized trials are characterized by random assignment of groups to study conditions, with outcome measurements taken on their members.

Examples:

- patients nested within clinics
- students nested within schools
- employees nested within work sites

7. Cluster or Group Randomized Trials

Outcomes measurements from members of the same group tend to be more alike than outcomes from different groups. (Observations not independent!)

The analysis should take into account the between-group variation (σ_b^2) in addition to the usual and the within-group (between-subject) variation (σ_w^2).

The *intraclass correlation* (ICC) quantifies the extent of between-group variation (or within group homogeneity)

$$\text{ICC} = \frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2}$$

7. Cluster or Group Randomized Trials

As the magnitude of the ICC increases, power will decrease assuming the sample size and other factors remain constant.

Let m = no. individuals sampled per group.

Design effect (DEFF) is the ratio of the no. subjects needed in a group-randomized trial vs. individual randomization.

$$\text{DEFF} = 1 + (m - 1)\text{ICC}$$

Note: When $m = 1$ or ICC $= 0$, the design effect is 1 and the two designs are the same.

7. Cluster or Group Randomized Trials

Denote the variance of the outcome within a group as $\widehat{\sigma}_w^2$ and assume this variance is same for all groups.

$$n = \frac{2\widehat{\sigma}_w^2[1 + (m-1)\text{ICC}]\left(t_{\alpha/2} + t_\beta\right)^2}{md^2}$$

$$m = \frac{2\widehat{\sigma}_w^2[1 + (m-1)\text{ICC}]\left(t_{\alpha/2} + t_\beta\right)^2}{nd^2}$$

$$= \frac{\widehat{\sigma}_w^2(1 - \text{ICC})}{nd^2/2\left(t_{\alpha/2} + t_\beta\right)^2 - \widehat{\sigma}_w^2(\text{ICC})}$$

The t-statistic has with df $= 2(n-1)$.

7. Cluster or Group Randomized Trials

How do I estimate ICC?

If data are available from previous study or pilot, the ICC can be estimated from a simple one-way ANOVA in which "group" is the only factor.

$$\text{ICC} = \frac{MS_b - MS_w}{MS_b + (m-1)MS_w}$$

where MS_b and MS_w are the between and within mean squares.

This formula is appropriate for both continuous and dichotomous endpoints.

7. Cluster or Group Randomized Trials

If data are not available, Gulliford *et al.* (*AJE* 1999;149:876-83); Hannan *et al.* (*AJE* 1994;5:88-95) provide ICC estimates for a range of risk factors and disease outcomes and geographic levels of clustering.

Otherwise, values between 0.01 and 0.05 can be used for sensitivity analysis. However, for large geographical areas, such as counties, ICCs will generally be less than 0.01.

7. Cluster or Group Randomized Trials

Example: Want to assess the effect of a behavioral intervention to lower cholesterol. Clinics are randomized to offer intensive dietary counseling or usual care. Find the number of clinics per condition to ensure 90% power to detect a difference of $d = 0.1$ mmol/l, using an $\alpha = 0.05$ level test. Assume 50 patients per clinic will be recruited.

Estimates of between- and within-practice variances were obtained from MRC thrombosis prevention trial.

$$\text{ICC} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_w^2 + \widehat{\sigma}_b^2} = \frac{.0046}{1.28 + .0046} = .00358$$

7. Cluster or Group Randomized Trials

$$\text{DEFF} = [1 + (50 - 1)(.00358)] = 1.175$$

This design requires 17.5% more subjects than if trial where individuals are randomized to treatment.

$$n = \frac{2(1.28)(1.175)(1.96 + 1.28)^2}{50(.1)^2} = 63$$

Note: If we chose to use 32 practices we would need 500 patients from each practice and the design effect would be 2.98. Thus the cluster design with 32 practices would require the total sample size to be tripled to maintain the same level of power.

7. Cluster or Group Randomized Trials

Example: Trial of Activity for Adolescent Girls (TAAG). Goal is to assess the effect of a school- and community-based intervention to prevent the decline in physical activity in adolescent girls. The number of schools per condition is fixed at 18. Find the number of girls per school needed to ensure 90% power to detect a difference of $d = 5$ *minutes of moderate-to-vigorous physical activity (MVPA)*, using an $\alpha = 0.05$ level test. Assume $\widehat{\sigma}^2_w = 400$.

7. Cluster or Group Randomized Trials

Between- and within-school mean squares, estimated with data from a prior study with $m = 58$ kids per school.

$$\text{ICC} = \frac{MS_b - MS_w}{MS_b + (m-1)MS_w} = \frac{11.31 - 3.54}{11.31 + (57)3.54} = .037$$

$$t_{.025,34} = 2.032,\ t_{.10,34} = 1.307$$

$$m = \frac{400(1 - .037)}{18(5^2)/2(2.032 + 1.307)^2 - 400(.037)} = 72$$

7. Cluster or Group Randomized Trials

Effect of adjusting for baseline value of the response:

$$\theta_w = \frac{\text{adjusted } \sigma_w^2}{\text{unadjusted } \sigma_w^2} = 0.9 \quad \text{and} \quad \theta_b = \frac{\text{adjusted } \sigma_b^2}{\text{unadjusted } \sigma_b^2} = .6$$

Can be approximated by $1 - R^2$, where R^2 is the squared correlation coefficient between baseline and follow-up.

$$m = \frac{\widehat{\sigma}_w^2(1 - \text{ICC})\theta_w}{nd^2/2\left(t_{\alpha/2} + t_\beta\right)^2 - \widehat{\sigma}_w^2(\text{ICC})\theta_b}$$

$$= \frac{400(1 - .037)(.9)}{18(5^2)/2(2.032 + 1.307)^2 - 400(.037)(.6)}$$

$$= 30$$